25X1

EDP STUDY

BR BIBLIOGRAPHIC PROJECT

30 November 1962

AUTOMATIC DATA PROCESSING STAFF

CENTRAL INTELLIGENCE AGENCY

Group 1
Exclude from automatic
downgrading and
declassification

Approved For Release 2003/04/29 : CIA-RDP84-00780R000200120058-6

Table of Contents

List of Illustrations

PART I.

SUMMARY

Soviet and Satellite scientific and technical publications serve as a major information source for scientific intelligence. The Air Force is currently supporting an index project covering selected open-literature materials of interest to aero-space intelligence. CIA's bibliographic effort, with input provided through an external contract with the Library of Congress, covers all the open Sovbloc scientific literature and represents the most complete index to this information. The author and organization files maintained by the Biographic Register provide information on the research being performed by Sovbloc scientists and organizations. These files are used not only by the intelligence community but by outside agencies such as the National Science Foundation and the National Academy of Sciences.

The principal problems which led to the initiation of this study are the large size and growth rate of the manual bibliographic files currently in use and the duplication of coverage between CIA and FTD. The major objectives of the study have been to determine whether electronic data processing techniques can be applied advantageously to this intelligence information activity and, if possible, to develop a coordinated Sovbloc literature exploitation program.

The study concludes that a cooperative effort between CIA and FTD in providing bibliographic control is both feasible and desirable. Not only is there complete duplication in the titles being covered in the

ILLEGIB

two projects but there exists a high degree of similarity in the information being extracted.

If it is useful.

It is recognized that the size and utilization of the bibliographic files make an automated system for storing and retrieving the data a necessity. It is recommended that, in preparation for the fully automated system, an interim system be implemented which would: (a) start capturing bibliographic input for machine processing as soon as possible; (b) use the input to generate, by computer, the 5 x 8 cards for manual filing into the author and organization files; (c) provide subject control for each entry using a magnetic tape file for storing the data; (d) begin experimentation with query techniques for performing subject search by computer.

Part II of the report reviews the history of bibliographic control both within and outside of CIA. Part III outlines the investigative procedures and discusses the major problems which had a bearing on the final recommendations. Parts IV and V contain detailed descriptions of the current CIA Bibliographic System and the Air Force's Project Cross Check. Part VI outlines the possible system configurations which were considered, concluding with a detailed description of the proposed interim system. Part VII discusses the problem of file conversion, recommending that action in this area be deferred until better criteria for conversion can be developed. Attention is also given to the manner in which an automated file could meet certain specialized user requirements, and the changes that would be introduced in the method of compiling the MIRA. Part VIII examines the input processing systems which

to capture as machine language for file conversion

might be set up with particular emphasis on CIA's relationship to Cross Check. A consolidated input effort under the single management authority of the Library of Congress is proposed. Finally, Part IX outlines the steps necessary to implement the proposed system.

The report also contains two appendices. Appendix A gathers together available statistics describing input and file utilization. Appendix B describes the proposed transliteration system for Russian text.

PART II.

BACKGROUND

The mission of the Bibliographic Project of the Biographic Register/OCR is to assemble and collate information derived from available Soviet Bloc scientific and technical literature in the form of comprehensive reference files organized specifically to facilitate intelligence research.

Intelligence has long recognized the value of exploiting this literature. Some ten years ago one CIA official declared that not only is the open literature the principal source of information regarding the research and development programs, organizations, and personalities of Soviet Bloc science, "in many instances [it] is proving to be the sole source of intelligence, and there is no immediate prospect of changing this circumstance through effective covert collection or otherwise."

What was true during the Stalin era in the USSR is still largely true today despite some relaxation of Soviet security. In the course of the DD/I system study now in progress, open literature has been cited consistently by ORR analysts as a prime intelligence source. The continued growth of specialized information collections devoted to the exploitation and control of the open literature is additional evidence of the importance attached to this kind of information as a source of intelligence.

A. Brief History of the Bibliographic Project

The Bibliographic Project dates from about September 1952 when the then Assistant Director for Scientific Intelligence, Dr. H. Marshall Chadwell, recommended the establishment within the Agency of a master file of Soviet scientific articles and books (particularly in the physical sciences) to be arranged by the name of the author and by the author's institutional affiliation. The argument for undertaking the project was based on the fact that efforts to derive information for intelligence purposes from existing source materials had been ineffective because nowhere was the information centrally indexed, collated, and filed in a suitable form to meet intelligence needs.

Author and subject indexes of commercial abstracting services, accessions lists, and similar research tools have been designed to serve the research scientist and technician who, almost invariably, seeks information on a world-wide basis. The intelligence analyst, on the other hand, typically treats and evaluates scientific events within a country as a unit. Moreover, these commercial indexes refer the researcher to bound volumes of abstracts, translated titles, and similar material. The latter in turn are in a form which makes it impractical, without reorganization of the data, to collect together physically and in the proper relationship the large bulk of information required for the kind of correlative analysis so important to the production of intelligence.

In order then to permit the intelligence officer to make more effective use of his available research and analysis time by providing him with the necessary raw data already collected together in file form, the Bibliographic Project was established and the responsibility for maintaining and retrieving the necessary information assigned to the Biographic Register.

In addition to author and organization control, it was recognized that a subject approach to the extracted data would be desirable. However, this would have required an additional index effort of a more difficult intellectual character, as well as a substantial amount of extra space to house the bibliographic collection, and the idea was therefore rejected.

To assemble the necessary data base as rapidly as possible, the entire file of carded abstracts (consisting of some 200,000 references) maintained at [redacted] was reproduced and arranged into the desired file sequences (author and organization) under an external contract. At the same time, arrangements were made with the Library of Congress (LC) to have 5 x 8 bibliographic reference cards prepared as a by-product of the effort devoted to the compilation of the Monthly List (now Index) of Russian Accessions. The method by which these cards are prepared is described in a later section of this report. Suffice it to say here that the MIRA has continued to be the major source of input to the project, which now contains in the neighborhood of 2,000,000 entries.

25X1

In 1959, a second agreement was signed with the Library of Congress--this time for the preparation of bibliographic cards on approximately 130 Satellite journals whose contents were being listed in another LC publication, the East European Accessions Index. As in the case of the Soviet material, only the scientific and technical publications were to be carded even though both the MIRA and EEAI cataloged titles in any subject field received by the Library of Congress and some 200-300 cooperating libraries in the United States.

Additional sources of input to the Bibliographic Project over the years have included such items as:

1. Copies of abstracts clipped from commercial abstract journals by Stork and others.
2. The entire science section of the 1949 Letopis Zhurnalnykh Statey (the Soviet bibliographic index to all periodical articles published in the USSR).
3. Abstracts, dissertation citations, and other bibliographic items translated from Sovbloc language materials by the Foreign Documents Division, CIA.
4. STEP Project abstracts of Bloc scientific and technical literature (described below).
5. Complete files of specialized subject interest compiled and maintained by individuals or groups in CIA's Office of Scientific Intelligence and elsewhere.

6. BR-prepared items including references to Bloc papers delivered at national and international meetings.

A major and continuing source of information on Bloc medical publications has been the National Library of Medicine (NLM). This agency has long furnished BR (via LC) with the author, title, translated title, pagination, reference source, subject headings, and biographic data for each article from Iron Curtain countries indexed in the Index Medicus. OCR currently reimburses the NLM for this service by furnishing sufficient funds to cover the rental cost of a Haloid Zerox 914 copier which is used to reproduce the NLM worksheets and other materials transmitted to LC.

B. Related CIA Projects

Despite the fact that the Bibliographic Project was designed to gain maximum coverage of the Bloc scientific and technical literature in order to serve as many intelligence needs as possible, it has not prevented other CIA units from sponsoring independent, specialized projects which have tended to duplicate the main effort. Some of these projects have been managed by a university or other non-governmental organization under a contract arrangement with an office in CIA. Others have been the private in-house efforts of individual CIA analysts, branches, or divisions. By no means are all such projects known even now--either to BR or to the study team. The following, however, are probably the most important open-literature index activities in terms of size and cost:

25X1 1. OSI [redacted] Project

This undertaking, formerly known as the Armed Forces Medical Library Literature Project, was established in 1954 under the terms of a contract between the Medicine Division/OSI and the

25X1 [redacted] The original scope of the project included:

a. Assembling files of abstracts and translated titles of published articles in the Soviet Bloc literature on the medical and related sciences, together with such other information as would permit collation of the material in any combination of author, subject, and institute relationship.

b. Provision of means for reproducing this material to whatever extent required to satisfy the analytical and research needs of the Medicine Division.

c. The preparation of summaries, reviews, statistical studies, etc., of the material in order to keep the Medicine Division informed as to developments in Soviet Bloc medical research.

When the project was originally proposed, it was criticized by the former AD/CR on the grounds that: (a) most of the Soviet medical literature was already covered in the Library of Congress' MIRA and in the AFML's Current List of Medical Literature (now Index Medicus); (b) author and institute information contained in most Soviet medical journals was already being filed centrally in BR's

Bibliographic Project. OSI argued, however, that although the MIRA and CLML carried subject indexes, these were essentially of the dictionary type and did not provide the degree of subject classification required to meet the Medicine Division's needs. (The Bibliographic Project itself, as pointed out above, did not include subject indexing.)

The issue was ultimately resolved by CIA's Project Review Committee in OSI's favor, and the project has since operated more or less along the lines originally proposed with the exception that no institute file was established since the Bibliographic Project could provide this control.

25X1 The [redacted] Project file now consists of approximately 530,000 cards organized by author within country (USSR or Satellites) and by subject. BR lends a significant amount of support to the undertaking since it furnishes the project with one set of all medical cards prepared by the Library of Congress. The only other important input to the file are photocopies of Excerpta Medica abstracts which the

25X1 [redacted] Project prepares itself. Extra copies of these abstracts are also sent to the Bibliographic Project where they are merged with other material in the collection.

Since the move of the National Library of Medicine to its new quarters in Bethesda, the project file and staff have been relocated in the Esso Building. The operation has been known as the

25X1 [redacted] Project since its absorption by OSI's [redacted] con- 25X1 tract. Contractual personnel make use of the file in preparing reports, bibliographies, and other studies on Bloc medical research.

The cost of the project in fiscal year 1962 amounted to $27,000.

2. OSI/USDA Literature Exploitation Project

This activity, unlike the [redacted] Project, was established about two years prior to the creation of the Bibliographic Project. Its original objectives apparently were to develop a file which would provide subject control of the Soviet Bloc literature in support of OSI intelligence production in the field of the plant sciences. With the passage of time, however, additional files were created, including an author, institute, and a source file. Ultimately, coverage was extended to include the field of veterinary medicine. 25X1

Like the [redacted] Project, this contract (which is physically located at the Department of Agriculture Library) duplicates much of the work of the Bibliographic Project since the latter has long included agronomy and veterinary medicine in its definition of the scientific fields to be covered in its program. However, again, the Bibliographic Project has provided no solution to the subject control problem. 25X1

BR does not support the D/A contract either directly or indirectly. Instead, the contracting organization acquires the bibliographic information it needs by reproducing the 3 x 5 work slips used by the D/A to prepare entries for its monthly Bibliography of Agriculture. These slips are reproduced in sufficient numbers for the project files (author, subject, journal, and organization). In addition, certain abstract journals are screened in a fashion similar to the [redacted] Project and pertinent abstracts are copied. Complete 25X1

translations are also prepared of selected titles of interest to OSI. These translations are title-indexed by FDD in its Consolidated Translation Survey. All veterinary science translations are also transmitted to BR where the names mentioned in the texts of the articles, together with their organization affiliations and fields of research, are indexed and punched into IBM cards.

The D/A project files now contain approximately 400,000 entries. In fiscal year 1962 the project's budget totaled $70,000, which included the cost of preparing translations.

3. Consolidated Translation Survey (CTS)

The CTS might also be considered to be performing a function comparable to that of the Bibliographic Project for it is a fact that the CTS file entries contain much of the same kind of information carried on the cards produced by the Library of Congress for BR and might well be used to answer some of the queries levied against the Bibliographic Project. The principal differences between the two operations, apart from the broader geographic and subject scope of the CTS, are as follows:

a. The CTS prepares bibliographic entries only on those Russian or Satellite books and articles that have been translated. (However, the number of such translations is, today, quite substantial.)

b. The CTS files its bibliographic records by author, subject, and source but not by institute affiliation of author.

The CTS file is in the form of 3 x 5 cards. Information entered on the cards includes: name of author, translated title (plus transliterated title in the case of books), language of the original, publication source, FDD-assigned serial number, and translation identifying information [redacted] number or other). The file was started in 1949 and FDD estimates its coverage is roughly fifty per cent scientific and technical. The total number of cards in the file is approximately 1,000,000, and it is growing at the rate of some 16,000 cards per month. Ten persons are employed in the project, which includes the maintenance of the file as well as the publication of the Consolidated Translation Survey. 25X1

C. Non-CIA Projects

There are many intelligence organizations outside CIA which also make some effort to index and control the scientific literature of the Bloc countries. Two of these exploitation programs have been of particular concern to those responsible for the direction and planning of the Bibliographic Project. They are the STEP Project and Project Cross Check, both sponsored by the Air Force.

1. STEP

The Scientific Technical Exploitation Project (STEP) was initiated in about 1958 with the aim of identifying and abstracting Soviet Bloc books and articles of interest to air intelligence.

The principal arguments used by the Air Force to justify the setting up of a new program rather than using existing bibliographic tools (including the MIRA and commercial abstract journals) were that:

(a) The time between publication and analyst acquisition of a translated abstract would be lessened greatly; (b) exploitation could be tailored to meet intelligence needs.

During its initial period of operation, STEP exploited some 200 scientific and technical periodicals, in addition to selected monographs, preparing abstracts of all the articles appearing in these publications. The abstracts were reproduced on 5 x 8 cards and included, in addition to the abstract of the text itself, the personality, affiliation, and other biographic data supplied the Bibliographic Project by the Library of Congress. Because the STEP abstracts were preferable to mere title references, BR arranged to obtain the entire STEP output. It has since replaced LC cards in the Bibliographic Project file with STEP cards whenever the latter have been available.

As the STEP Project gained momentum, it soon became obvious that it was a wasteful duplication of effort to have the Library of Congress index, for CIA's Bibliographic Project, the same titles covered by STEP. Accordingly, in about 1959 all titles exploited by the Air Force program were dropped from BR's Library of Congress contract and replaced by selected East European materials. Less than a year later, however, because of budget restrictions and criticisms concerning the value of some of the items abstracted, Air Force officials required STEP to drop its cover-to-cover program and become more selective in choosing the items to be abstracted. This, naturally, prevented BR from relying further on the program since it

was impossible to anticipate what titles STEP would choose to process. All titles, therefore, previously dropped from the LC list were reactivated.

STEP abstracts are still being produced and, indeed, the scope of coverage has been widened to include some 500 journals. The cards themselves are being filed not only by BR, but by the Aerospace Information Division (AID) of the Library of Congress and, on a selective basis, by [redacted] The last-named organization also examines the Russian bibliographic journals Knizhnaya Letopis (Book Index), Letopis Zhurnalnykh Statey (Index to Periodical Articles), and the various sections of the Referativnyy Zhurnal (Abstract Journal) for references to publications not available outside the Iron Curtain. Items found are added to [redacted] own open-literature collection. To further complicate the literature exploitation and service picture, the study team was told by sources at the Air Force's Foreign Technology Division (FTD) that OSI is paying [redacted] some $40,000 per year for information support.

25X1

25X1

25X1

2. Cross Check

This Air Force literature exploitation effort more directly duplicates the Bibliographic Project than any other, and was one of the major reasons for the initiation of this study. Begun in 1961, "the primary aim of the project is the compilation of:

a. Lists of scientific and technical personnel of organizations involved in research, development, and production work.

b. Lists of organizational and/or subject associations of scientific and technical personalities."

Both the aims of the project as well as the source documents exploited are identical to those of the Bibliographic Project. All of the open-source materials selected for indexing by Cross Check are included in BR's program. Similarly, the items to be extracted by Cross Check (i.e., name of personality, identification of source material, name of organization with which the personality is associated, indication of field of competence of personality, etc.) are, with minor exceptions, the same as those of the Bibliographic Project. The only significant differences between the programs are that:

a. Cross Check is extracting names mentioned in the texts of the journal articles, as well as authors.

b. The Bibliographic Project indexes more titles because its scientific subject interests are more comprehensive.

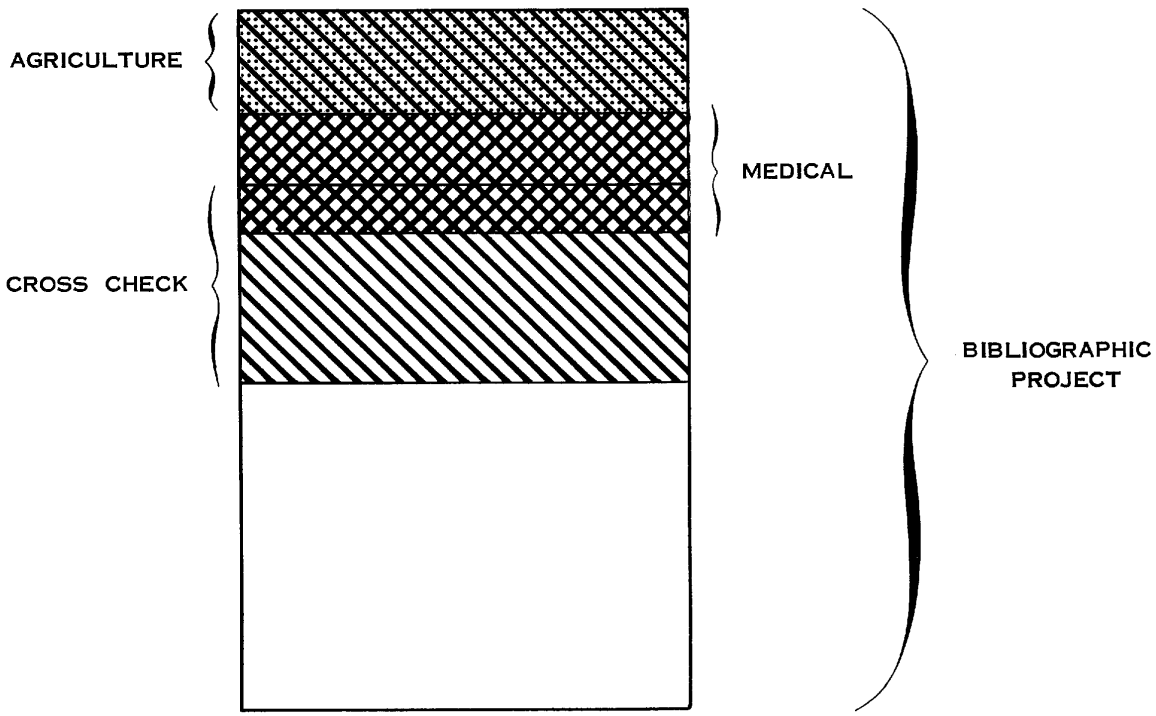
c. Subject searches are a planned facet of the Cross Check retrieval program.

d. Cross Check is a computer-supported system while the Bibliographic Project is a manual operation in its entirety.

Cross Check's input is performed by personnel located at AID. These individuals collect and control the source materials, scan the text of these materials, extract the required data, prepare worksheets with the information arranged in a specified format, and con-

vert the worksheet entries into computer language through the use of punched paper tape. The completed tape is forwarded to the sponsoring agency to be fed into the computer.

Cross Check began its coverage with the first issues of journals published in 1961. It hopes to have processed into tape all of the 1961 and half of the 1962 issues of these journals by the fall of this year. The project currently employs 25 input people full-time at AID, plus part-time programmers and monitoring personnel at FTD. The University of Dayton is also punching some of the Cross Check data under contract with FTD.



S & T LITERATURE COVERAGE BY PROJECT

FIGURE 1

PART III.

STUDY APPROACH

The study phase of the team's investigation of the Bibliographic Project was directed toward gathering information in the following areas:

a. Material covered for inclusion in the Bibliographic Project files and for use in producing the MIRA.

b. Current procedures for generating input to the files and to the MIRA.

c. Structure of the files.

d. Purpose of the files and the ways in which they are used.

e. Other USSR and Satellite scientific and technical bibliographic projects within CIA.

f. Major USSR and Satellite scientific and technical bibliographic projects outside of CIA.

A. Methodology

At the start of the study, a visit was made to LC to determine the procedures being used in preparing inputs to both the Bibliographic Project files and to the journal, MIRA. For each of the input categories (Russian monographs, Russian periodicals, Russian medical literature, East European monographs, East European periodicals, East European medical literature) all of the processing steps performed by LC were noted.

Having determined the procedures used in preparing input to the Bibliographic Project, a visit was made to the Support Branch of the

Biographic Register to continue the tracing of the input as it is processed for inclusion in the card files. Here the composition and content of the files were examined and the exact steps for preparing new card inputs were recorded.

The interrelationships of the Bibliographic Project files and the other files maintained by the Biographic Register were then considered. Informal discussions were held with analysts within the USSR Section of BR to determine how they used the files and to obtain their suggestions for improving the files.

Information on the OSI [redacted] Project was obtained from Dr. [redacted] of OSI. He evidenced great interest in the study and was most optimistic that the OSI contract for a medical file could be terminated if BR were able to provide better subject control over the medical literature. 25X1 25X1

The study of the Air Force's activities in open-literature indexing began with a visit to AID to determine the material covered for Project Cross Check and the procedures used in preparing the inputs, including examination of the exact items of information provided in each entry. Subsequently, a trip was made to FTD where discussions were held with the sponsor regarding the purpose of the project, the uses to be made of the data, and the techniques which would be used to manipulate the data. It was determined that there was significant overlap between input preparation for the Bibliographic Project and for Cross Check. The people at FTD expressed interest in developing a cooperative input preparation effort and agreed to

consider changes which would make their data compatible with the needs of CIA.

A follow-up meeting was then held at CIA with personnel from FTD and AID at which a detailed description of the Bibliographic Project, including discussion of the MIRA, was presented. On this occasion FTD requested a memorandum delineating the input requirements of CIA and, in particular, pointing out what we considered to be the areas of incompatibility along with suggestions for eliminating the difficulty. This was done and, subsequently, another meeting of CIA, FTD, and AID was held to discuss in detail all suggested modifications. Again, interest was expressed by the FTD representatives in eliminating all conflicts so that a cooperative effort might be established.

Based upon the visits and interviews, a system design was developed which is presented in this paper.

B. Design Considerations

The basic assumption underlying the system design for the Bibliographic Project is that the bibliographic files, their coverage, content, and responsiveness, must, first and foremost, satisfy the needs of the intelligence analyst. Keeping in mind this major premise, two additional assumptions were accepted: (a) that coordination with the Air Force in preparing input to the files is desirable; (b) that the publication, Monthly Index to Russian Accessions, is a valuable tool at least for the academic community.

In the course of developing the recommended system design, many problems were considered and decisions made which had a bearing on the final recommendation. These design elements will be summarized here.

1. Timeliness of Information

One of the first considerations was that of providing information to the analyst more rapidly than is currently possible. Because of the time lag which exists between the date of publication of research activity and the discovery and evaluation of this research by an intelligence analyst, it is necessary to make some provision for the possibility of error in estimating the current state-of-the-art in a given scientific field. To superimpose on it a delay factor of from six months to one year in making available to the analyst references to the published material is a serious hindrance to the development of current intelligence. The introduction of automation as proposed in this paper would reduce this delay factor.

2. Data Sources

Another problem was that of specifying the components of the data base. The advisability of exploiting the Russian abstract journals and the published indexes to books and periodicals was considered. It was finally decided that coverage of these sources could not replace coverage of the books and periodicals themselves, because: (a) no organization affiliation for personalities is provided in them; (b) it would mean relying on the Russians to decide what elements of their scientific and technical literature are of primary

interest to U.S. intelligence; (c) it would mean an added delay in making information available because of the time required for the Russians to cover their material and publish it in abstract journal or index form.

References generated within CIA (CIA Library accession cards, FDD references, and items published in intelligence reports), international conference documents, and STEP cards have been eliminated, at least for the time being, from the data base. All of these materials duplicate, to a large extent, LC coverage. Since a complete STEP file is maintained in Washington at AID, the abstracts can be readily obtained whenever needed.

In regard to indexing monographs which are compilations with individual chapter authors, it was decided that, if cost permits, coverage of the individual chapters would be a worthwhile addition to the file. In effect, such monographs are much like periodicals and coverage by chapter is as valuable as article coverage within journal.

3. Elements of the Index Entry

In specifying the content of each index entry, several alternatives were considered. First, a decision was made on name coverage. It is felt that names in the text, in footnotes, and in the bibliography should not be extracted at this time; the added cost is not warranted since the intelligence information on these individuals (field of interest, organization affiliation, etc.) is normally obtainable from direct references to them as authors.

Two new elements of information are recommended for incorporation in the basic entry. They are a Field of Interest code and a Nationality code for each individual. Both codes are provided in anticipation of the fully automated system. The Nationality code permits a single bibliographic file while allowing for queries which specify nationality. The USSR and the Satellite countries will each have a unique code. Other nationalities will be coded as "all other." It is felt that the number of entries in the "all other" category will be small enough to permit manual selection, from the machine print-out, of the nationality group desired. In particular this category will be most useful in providing access to Chinese scientists publishing in the USSR or Eastern Europe.

The Field of Interest code is used to differentiate individuals with the same name working in different fields thus reducing the false drop rate on retrieval. Further study must be given to the degree of specificity needed here.

4. Indexing Rules

One of the indexing problems is that of representing organizations. In current LC processing, organization affiliation is carried on the bibliographic cards as a transliteration of the full organization name. In the Cross Check system, organization affiliation is represented by an alphabetic code with a distinct code for every organization to the smallest component. For machine processing, it is not reasonable to use the full organization name because minor variations of the same name will make machine match impossible;

the probability of a single character error which will again make machine match impossible, is much greater in transcribing or in typing a long title than in a short mnemonic code. Moreover, many organization titles, when given with Academy affiliation, etc., are very long and the use of a code would significantly shorten the machine record.

After deciding on the use of a code, it was necessary to determine how far down the organization hierarchy to code. As a general rule, it is planned that coding will be carried to the level of institutes but, in special cases, more or less detail may be provided. In order, however, to allow for more specific coding in the future, for all organizations which are cited in greater detail in the source document, the transliteration of the full title will be included in the record.

Another of the problems considered was that of how to provide subject control. Cross Check is indexing articles and monographs using keywords; the words contained in the title form the basic index and additional words are extracted from the text to more fully describe content where necessary. It is felt that, for the material of interest to CIA, keywords should form a basic part of the index because: (a) such an index may be easier and cheaper to apply than a classification index; (b) it can provide more detailed indexing; and (c) it makes the input compatible with the Cross Check product. However, indexing with predefined subject categories must also be considered and, before designing in detail the final indexing

system, existing systems for indexing scientific and technical literature should be examined and experimentation performed to develop the type of index which would best serve the intelligence analyst.

5. Transliteration

In preparing for an automated system in which searching and matching will be one of the major operations, it is necessary to achieve compatibility in all inputs. Moreover, as machine processing of text increases throughout the Community, it is imperative that there be compatibility between organizations as well as within projects. More specifically, if cooperation is to be achieved between Cross Check and the Bibliographic Project, these two inputs must be compatible. But even within the LC-produced data there exists an amazing conglomeration of transliteration systems which must be coordinated. In response to this need, a system for transliteration from Russian to the Latin alphabet, which is a minor variation on the BGN system, was developed (Appendix B). Its use not only standardizes the transliteration but provides text from which all of the original Cyrillic characters may be unambiguously determined. At the same time, it provides a representation which is easily read by the user. This, however, solves only a portion of the total problem. Under the current system, the same name, in different languages, will transliterate differently so that machine search for a given individual necessitates the representation of the name as it would appear in all of the different languages being covered. This is obviously undesirable. At least for the languages within the USSR, some

standardization must be provided. A system must, therefore, be specified for all of the non-Russian languages of the USSR and for each of the languages of Eastern Europe. It is desirable to devise a method which will provide for unambiguous representation of each character in the original without resorting to the use of diacritics which, at present, are not available as standard equipment on computer printers.

PART IV.

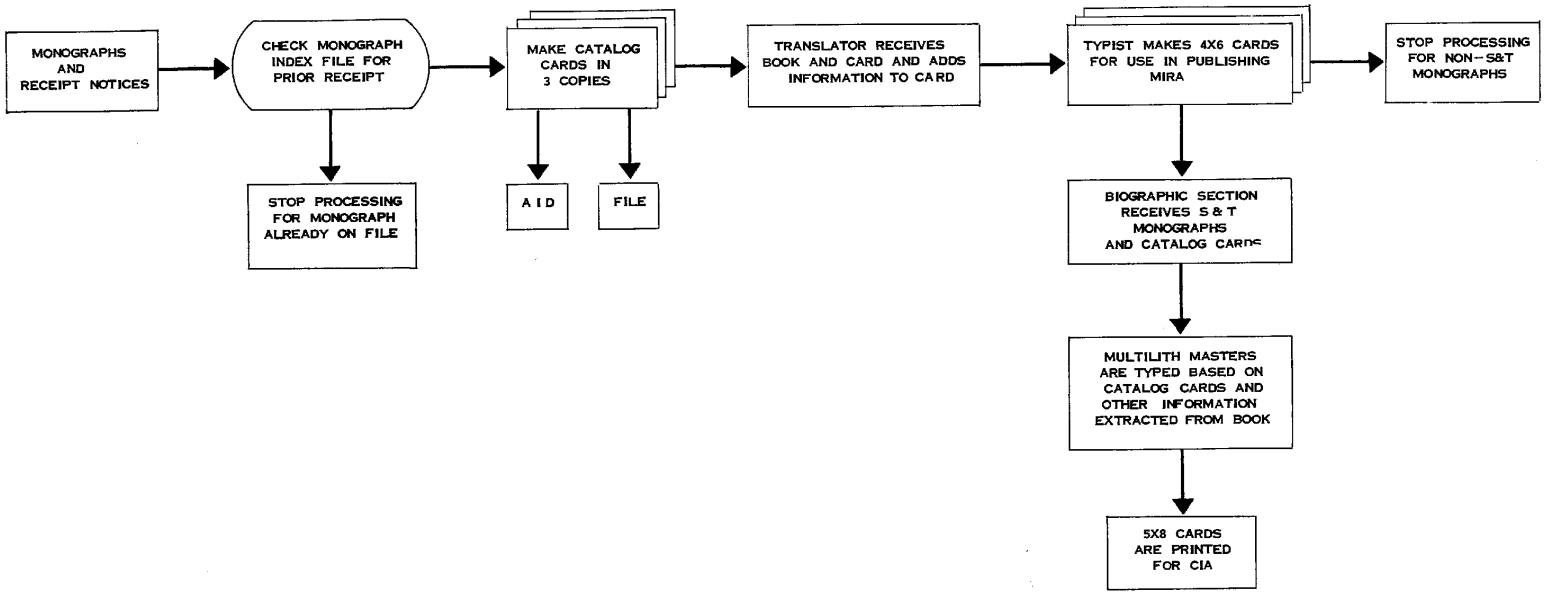
PRESENT BIBLIOGRAPHIC SYSTEM

The majority of the input for the Bibliographic Project files maintained by the Biographic Register is produced by the Library of Congress. All Russian monographs and periodicals and all East European periodicals accessioned by the Library of Congress or by one of the cooperating libraries are covered. Information extracted from medical journals received and processed by the National Library of Medicine is forwarded to the Library of Congress for inclusion in the produce sent to CIA.

A. Input Processing at LC

1. Russian Monographs

Monographs, or notification of receipt of monographs by cooperating libraries, are received at LC where the titles are checked against the title index for monographs to determine whether the book has already been processed. If it has not, three copies of a file catalog card are produced containing transliterated title, author, and publication information. One copy is entered into the file; a second copy is forwarded to AID; the third, along with the monograph, is routed to a translator based on the subject of the book. The translator adds the translated title and subject headings to the file catalog card. He selects the subject headings from the LC index entitled Subject Headings. At this point, a typist produces 4 x 6 cards based on the information on the file catalog card for



PROCESSING RUSSIAN MONOGRAPHS

AT LIBRARY OF CONGRESS

FIGURE 2

SECRET

use in producing the publication, Monthly Index to Russian Accessions (Figure 3). A separate card is typed for each subject category assigned to the monograph plus one additional card for use in preparing Part A of the journal. The monograph and its catalog card are next routed to the Biographic Section where all scientific and technical material is selected for further processing. One person is then responsible for producing Multilith masters based on the information contained on the catalog card plus additional information extracted from the monograph (Figure 4). The following information is added to that already extracted:

a. additional authors mentioned on the title page or in the table of contents of the monograph.

b. names of editors, technical editors, and author's superiors.

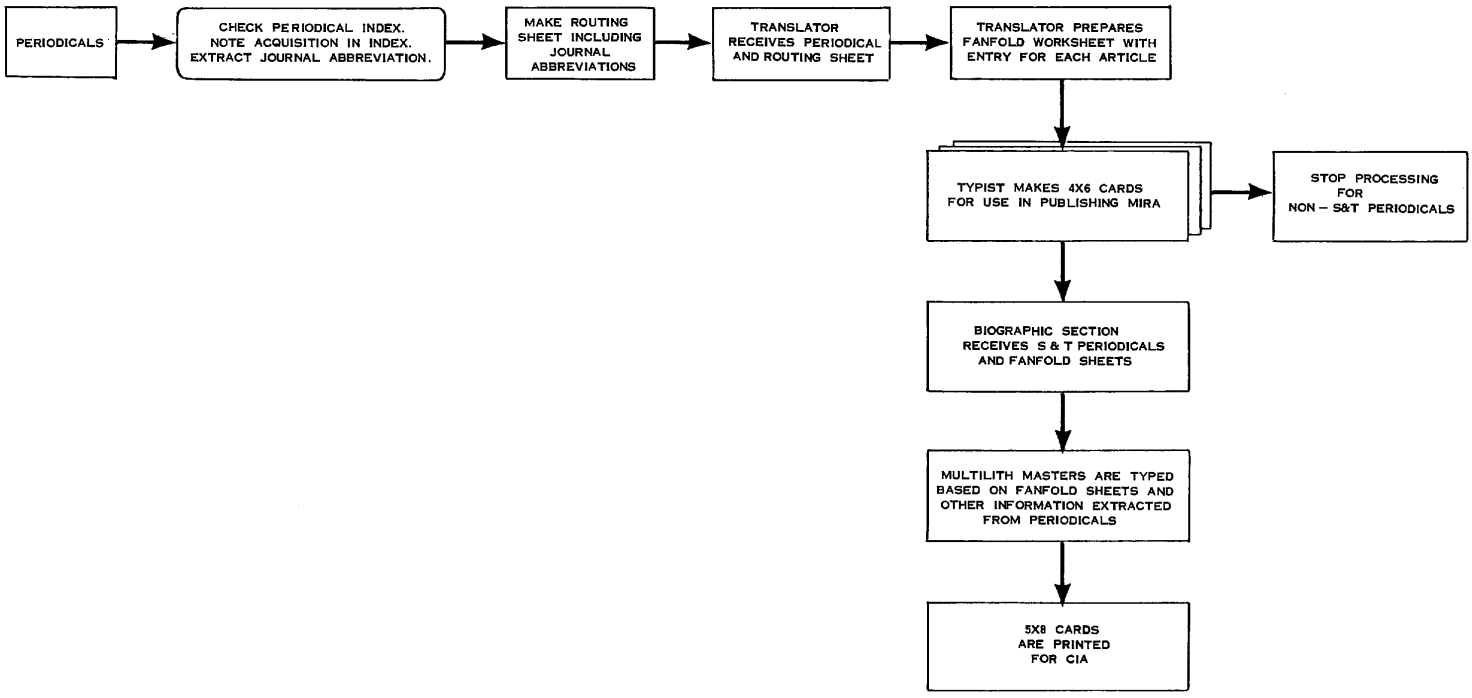
c. titles and organization affiliations for all personalities.

Full names are extracted when available. BGN transliteration is used for all names of individuals and organizations extracted.

2. Russian Periodicals

As each periodical is received at LC, its accession is noted in the periodical index maintained on a rotary card file. At the same time the official journal abbreviation is extracted from the file and noted on a routing sheet which, together with the periodical, is sent to a translator. The translator prepares, on fanfold paper, a typewritten entry for each article in the periodical. Each entry

Next 1 Page(s) In Document Exempt



PROCESSING RUSSIAN PERIODICALS AT LIBRARY OF CONGRESS

FIGURE 5

contains the English translation of the title of the article, a transliteration of author names according to the LC system, subject headings, and the page number of the first page of the article. The name of the journal in which the articles appeared and other journal identification information are typed once at the start of the series of pages covering the articles for that journal. A typist then produces 4 x 6 cards from the fanfold sheets for use in producing the MIRA (Figure 6). For each article covered, one 4 x 6 card is typed for each subject heading assigned to the article. An entry is made in the periodical index and on the fanfold sheets specifying the issue of the MIRA in which the reference will appear. The periodical and its associated fanfold sheets are then sent to the Biographic Section where scientific and technical material is selected for further processing. To the information contained on the fanfold sheets received with the periodical, two analysts add the following:

a. The BGN transliteration of author names with the names in full, if possible. Names of author's superiors, if available, are also added to the worksheet.

b. Title and organization affiliation for each author (Figure 7).

Multilith masters are then typed from which 5 x 8 cards are made for CIA. A serious backlog exists in the typing of Multilith masters. As much as six months processing is waiting for typing.

Some remarks the propagation of electromagnetic waves in an anisotoropic dispersive medium with summary in English. B.N. Gershman, V.L. Ginzburg. Izv. vys. ucheb. zav.; radiofiz. 5 no. 1:31-46 '62

ELECTROMAGNETIC WAVES

Vertical focusing of magnetic analyzers of mass spectrometers. R.N. Gaill'. Zhur. tekh. fiz. 32 no. 4:402-405

MASS SPECTROMETRY

Vavilov-Cherenkov phenomenon and the intensification of ultrasonic waves (from "Scientific American," 205, No. 5, 1961) Priroda 51 no. 5:117-118 My '62.

ULTRASONIC WAVES

Phytopatholgic effectiveness of some ways of presowing treatment of wheat seeds. Z.P. Kachalova. Dokl. TSKHA no. 41:115-120 '59.

WHEAT

SAMPLE MIRA CARDS

FIGURE 6

Ogneuporty 26 no. 9 '61

jgb - 1

Refractory Kinas material. P.N. D'iachkov and others. 394-398

REFRACTORY CONCRETE

D'yachkov, P.N.
Purgin, A.K.
Bol'shakov, I.P.
Gubko, I.T.
Kostomarvov, M.I.
Sizov, I.D.
1. Vostochnyy institut ogneuporov (for D'yachkov, purgin, Bol'shakov).
2. Pervoural'skiy dinasevyy zavod (for Gubko, Kostomarov. Sizov).

Formation of mullite in short-prism, isometric form and its effect on the refractoriness and deterioration of fire clay articles. K.K. Strelov. T.F. Raichenko. 431-436

1. MULLITE

2. Fire Clay

Strelov, K.K.
Raychenko, T.F.

1. Vostochnyy institut ogneuporov.

FANFOLD FOR RUSSIAN ARTICLES

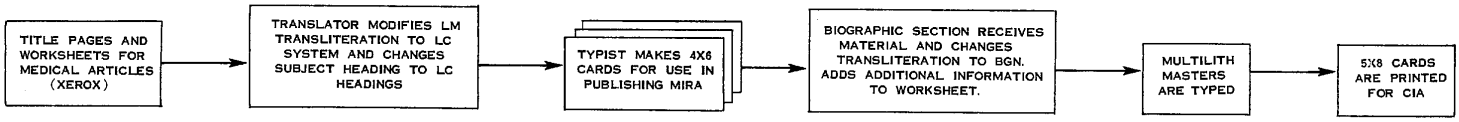
FIGURE 7

3. Russian Medical Periodicals from NLM

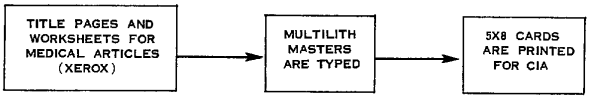
The National Library of Medicine sends to the Library of Congress Xeroxed material covering articles from Russian medical journals. Contained in the Xeroxed material for each journal is a copy of the title page of the journal, a copy of the first or title page for each article in the journal, and a copy of the NLM worksheet for each article. The worksheet contains the translation of the article title, the transliteration of author name using the NLM transliteration system, and a listing of the NLM subject headings for the article. For preparing input to MIRA, the NLM transliteration is modified to agree with the LC system and the subject headings are altered to correspond to the LC subject heading index. A 5 x 8 card is typed for each subject category assigned to the article. These cards are used in the publication of the MIRA. The Xeroxed material is then sent to the Biographic Section where the author's name is transliterated again in the BGN system. Titles and organization affiliations for the authors are added to the worksheets. Finally, Multilith masters are prepared containing the required information and 5 x 8 cards are produced for CIA.

4. East European Medical Periodicals from NLM

Xeroxed material covering articles from East European medical journals, similar in content to that covering Russian medical journals, is received from the Library of Medicine (Figure 9). It is sent directly to the Biographic Section. Multilith masters are typed directly from the worksheets as received from NLM with no



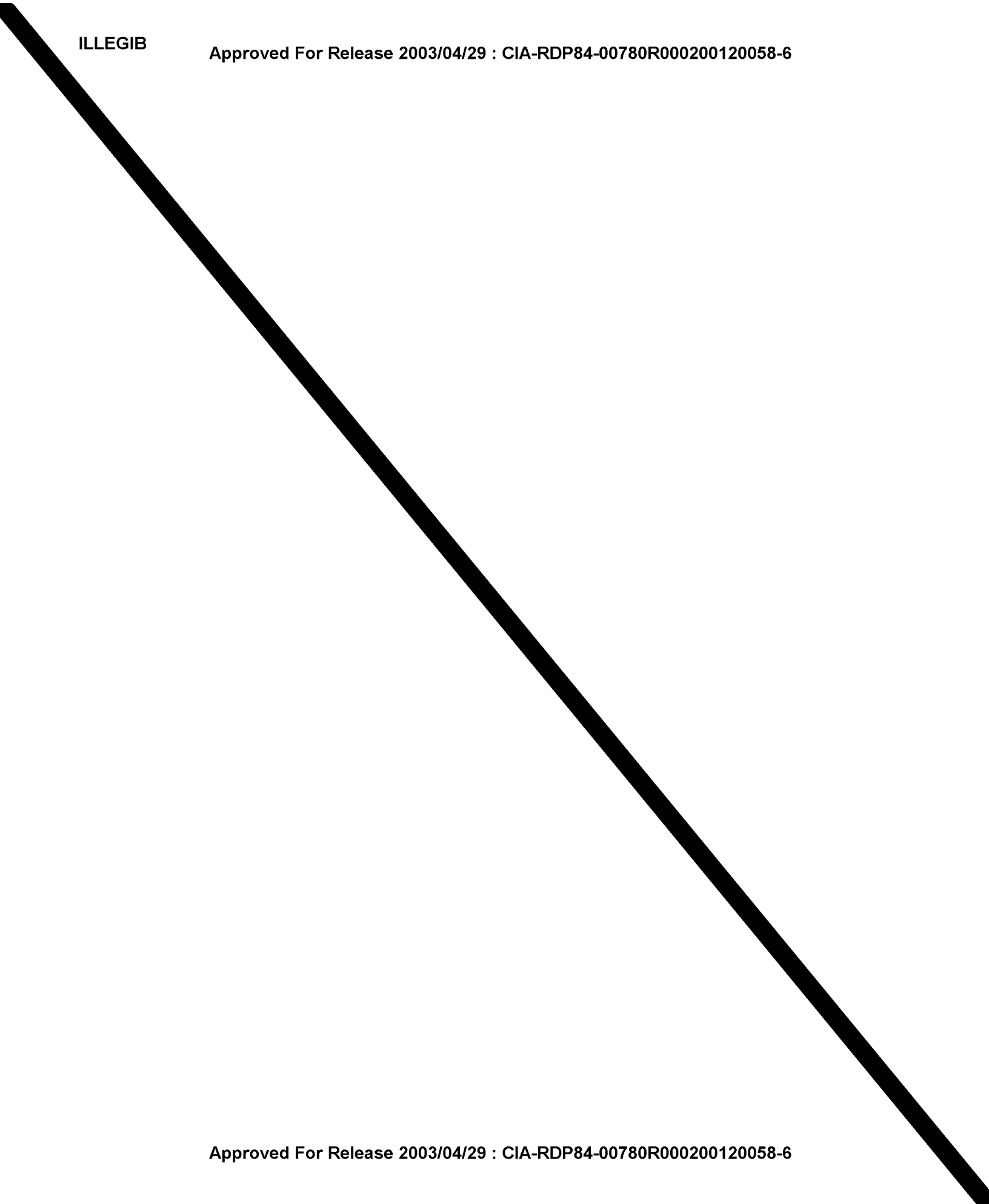
PROCESSING RUSSIAN MEDICAL JOURNALS AT LIBRARY OF CONGRESS



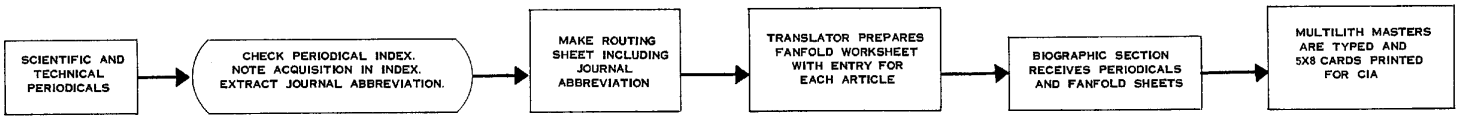
PROCESSING EASTERN EUROPEAN MEDICAL JOURNALS AT LIBRARY OF CONGRESS

FIGURE 8

SECRET



Approved For Release 2003/04/29 : CIA-RDP84-00780R000200120058-6



PROCESSING EASTERN EUROPEAN PERIODICALS AT LIBRARY OF CONGRESS

FIGURE 10

changes to transliteration or subject headings. 5 x 8 cards are produced from the Multilith masters for CIA.

All cards covering medical references, both Russian and East European, are duplicated and one set, with one additional card, for each reference is sent to OSI's [redacted] Project. A set of cards consists of a card for each author and organization named in the reference. The [redacted] Project files the cards by author within country and by subject, substituting its own subject categories for those provided on the cards.

25X1

25X1

5. Other East European Periodicals

When East European periodicals are received, accessions are noted in the periodical index and a routing slip is prepared carrying the official abbreviation for the journal. The periodical is then sent to a translator based upon the language in which it is written. These translators, of which there are six, are physically located with the Russian translators. The translator prepares fanfold sheets with an entry for each article in the journal. Since the personnel in the Biographic Section do not know the East European languages, the translator extracts both the information normally extracted for the MIRA processing of Russian journals and the information usually appended by the Biographic Section. The completed fanfold sheets (Figure 11) are sent to the Biographic Section where they are cursorily edited and then typed on Multilith masters. 5 x 8 cards are prepared for CIA from the mats.

Aplikace mat 5 no. 6 '60

A contribution to the study of singular points in photos elasticity. 401-411

(Photoelasticity)

Svecova, Hana

1. Author's address: Matematicky ustav. Ceskoslovenska akademie ved. Praha-Nove Mesto, Zitna 25.

A numerical calculation of quasi-stationary solution of heat conduction equation. 412-441

(1. Heat
2. Equations)

Vitasek, Emil

1. Author's address: Matematicky ustav. Praha-Nove Mesto, Zitna 25.

Guldberg-Waage transformation in homogenous reactions. 442-452

(Transformations (Mathematics)

Lansky, Milos, dr.

1. Author's address: Katedra matematiky a fysiky pri Pedagogickem institutu, Karlovy Vary, trida Jednotnych odboru 11.

FANFOLD FOR EASTERN EUROPEAN ARTICLES

FIGURE 11

B. Publication of the MIRA

As the 4 x 6 cards are typed for use in the MIRA, they are sorted and interfiled with those cards waiting for publication. At the end of the month, the full set of cards is shingled up and photographed and the journal is printed. The cards are then destroyed.

The journal is published in three sections. The first is a list of all monographs in alphabetic order by subject and by author within subject. The second is a list of all periodicals whose articles are referenced in the issue. The final section consists of a subject index to monographs and periodical articles; the entries within each subject category are in order by title.

C. Bibliographic Input Processing at CIA

In addition to the primary source of input produced by the Library of Congress, Russian dissertations and selected Satellite periodicals are monitored by the Foreign Documents Division of CIA. The CIA Library also provides information regarding its accessions of Russian and Eastern European scientific and technical monographs, and the Biographic Register generates input to the bibliographic files based on bibliographic references in intelligence reports and international conference documents. 5 x 8 cards are prepared for all of these references.

The Air Force STEP cards are another input to the Bibliographic File. These cards provide coverage of selected scientific and technical articles from Russian and Satellite periodicals. In addition to the information normally contained on the CIA bibliographic cards,

the STEP cards carry an English language abstract. STEP title coverage duplicates title coverage by LC for CIA. Because of the abstract which they provide, they are used to replace the LC cards.

Finally, over the years, information from ad hoc exploitation programs, such as the [redacted] abstract cards, as well as files compiled by other offices and agencies, have been incorporated in the collection. 25X1

D. Maintaining the Bibliographic Files

When the 5 x 8 cards are received by BR, personality names are checked for spelling validity. The cards are then divided into three major decks: cards for the author file, cards for the USSR organization file, and cards for the Satellite organization file. The Satellite organization cards are sent to the Satellite Section for sorting and filing. USSR organization cards remain in the Support Branch where they are sorted and filed. For author cards, the sort key is underlined and the cards are sent to the central CIA clerical pool (Interim Assignment Branch) for sorting. When they are returned from the pool the Support Branch interfiles them. In the filing process, references which duplicate cards already on file may be encountered. In that case, the new reference is discarded unless it provides more complete information than the card currently on file (e.g., an abstract) in which event the new card replaces the old.

In the course of checking new cards when they first arrive, copies of references to published biographies are pulled for inclusion in the dossier or biographic card ("consolidated") file and a

request is sent to FDD, based on the reference, asking that the biography be translated.

E. Querying the Bibliographic Files

Queries in which the interrogation criterion is either author name or organization may be answered directly by entering the appropriate file. Subject category access to the information on file is obtained by a two-step procedure. First, the searcher must attempt to determine all organizations which might be doing research in the desired area. He must then hand-search the rather extensive set of cards associated with these organizations in the file, examining subject headings and titles contained thereon to determine those references which might be relevant.

In most instances, when querying the files, the analyst removes the cards which are of interest thus necessitating the refiling of these cards.

PART V.

PROJECT CROSS CHECK SYSTEM

The Aerospace Information Division (AID) of the Library of Congress processes input for the Air Force in support of Project Cross Check. Approximately 220 scientific and technical Russian journals are covered. In addition some data are extracted from selected newspapers and secondary sources such as the Russian abstract journals. In the case of the secondary sources, information is selected if the subject is of interest and if the abstract covers an article contained in a journal other than those normally processed in the system.

A. Input Processing at AID

For each article to be covered, a transcript sheet is filled out by the translator. The following information is typed on the form:

1. Translation of article title. (If the reference is from an abstract journal, this is followed by the name of the journal in which the article itself appeared.)
2. Keywords descriptive of the article content to supplement the keywords contained in the title. (A maximum of 25 keywords may be entered by the translator. If the article is scheduled to be abstracted by STEP, STEPA is included as a keyword for the article.)
3. Date on which the article is being processed.
4. Numeric identification code for the journal. (In the case of references extracted from abstract journals, the code for the abstract journal itself is used.)

5. Year of publication, volume number, and issue number of the journal.

6. Page number of the first page of the article.

A set of information is then included for each personality of interest who is associated with the article. The latter include: authors; names in the bibliography associated with a Trudy, dissertation, patent reference, or an article published in a source which is not accessible in this country; names in footnotes; and names in the text. Names which occur in both the text and the bibliography or both the text and the footnotes are indexed based only on their appearance in the bibliography or the footnotes. For each personality, the following information is extracted:

1. Personality name.

2. Page number on which the name appears. (In the case of authors, the page number used is that of the first page of the article in all cases.)

3. Professional status or title of the individual.

4. Code for relationship of the individual to the article (e.g., author, name in text, etc.).

5. Mnemonic code for organization affiliation of the individual.

6. Geographic location of the individual. (This is used when no organization is specified and a location for the individual is available.)

Information Field	Bibliographic Project	Project Cross Check
Accession Number	No	Identification number assigned to each article or monograph.
Title	Translated	Translated
Subject Coding	LC subject headings	Keywords extracted from text
Processing Date	No	Date worksheet is prepared
Journal Identification	Abbreviation of journal name	Numeric code
Publication Date	Month and year	Year
Volume	Yes	Yes
Issue	Yes	Yes
Pagination	Number of first and last page	Number of first page
Personality Name	Names of authors, editors, and author's superiors.	Names of authors, editors, author's superiors and names in footnotes, bibliography, and in the text.
	Transliteration of professional title	Transliteration of professional title
Organization Affiliation	Transliteration of full organization name	Alphabetic abbreviation
Relationship	Location of personality name on card indicates whether author or author's superiors. Name in brackets indicates deceased.	Code indicating relationship of personality to organization or source material.
Number of Page Where Name Appears	No	Yes
Location	No	Yes
MIRA Reference	Volume and issue number of MIRA journal in which item will appear.	No

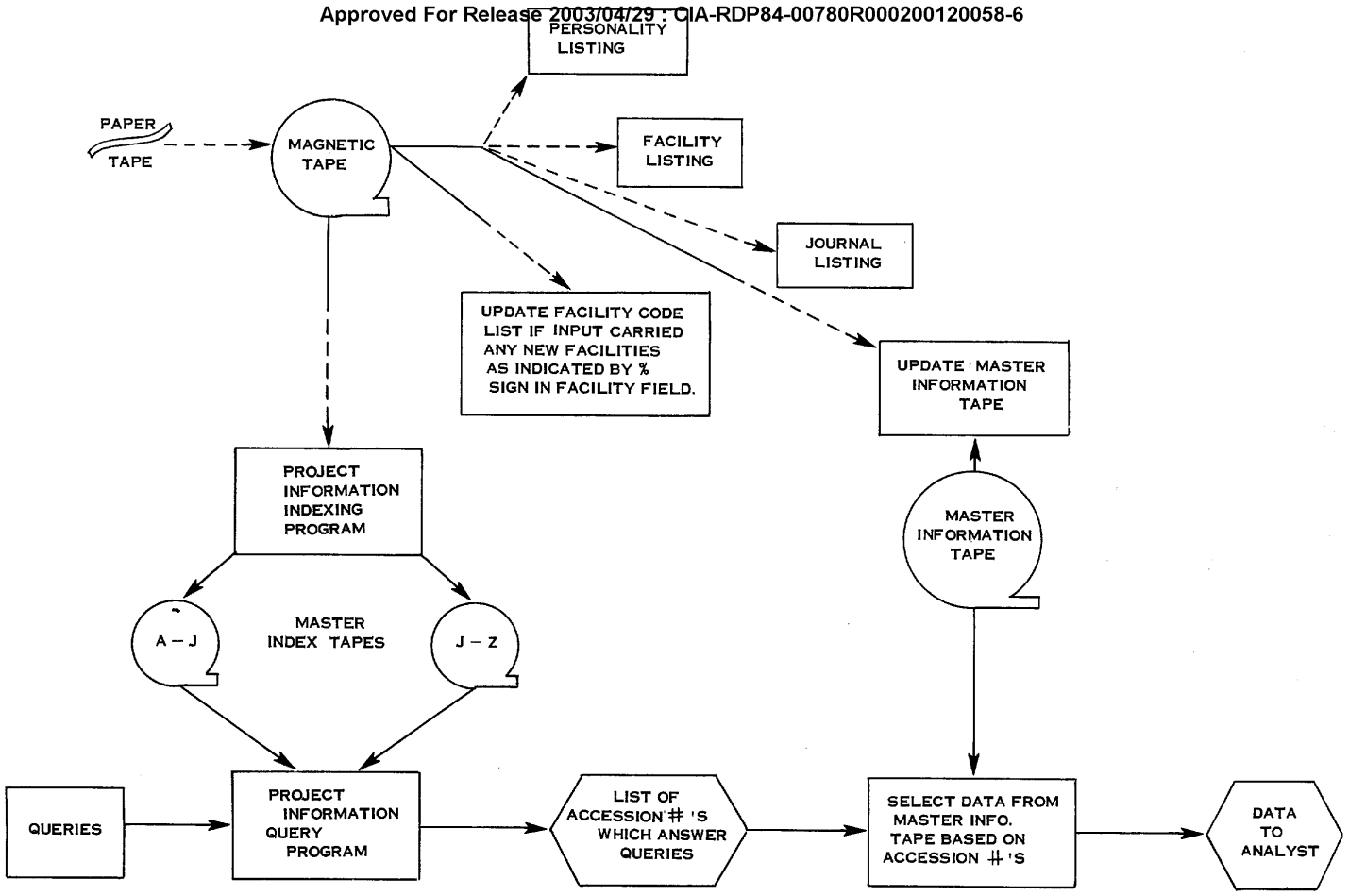
COMPARISON OF INFORMATION EXTRACTED BY BIBLIOGRAPHIC PROJECT AND CROSS CHECK
FIGURE 12

The completed transcript sheets are routed to the Flexowriter section where they are assigned sequential accession numbers. The information is then typed using Flexowriters to produce both a hard-copy output and paper tape.

B. FTD Processing

At FTD the paper tape information is converted to magnetic tape for use in updating the machine readable data files. Periodically, three listings of the data are produced: a listing by personality, a listing by organization or facility, and a listing by journal.

In addition to these listings, magnetic tape files will be maintained for use in answering queries. These files will consist of a Master Information Tape which will contain all of the data records, in full, and a Master Index Tape which will contain, for each element which can be used as a selection keyword, a list of all pertinent records on the Master Information Tape. The elements which may be used as selection keywords are: all words in the title field, excluding common words; each word in the keyword field; the journal reference code; year of publication; volume number; issue number; personality name; organization code; and location. Any logical combination of the above keyword elements (not including negation) may be used to query the Master Information Tape. FTD's Project Information query program will be used to interrogate the files.



PROJECT CROSS CHECK PROCESSING FLOW

FIGURE 13

PART VI.

POSSIBLE SYSTEM CONFIGURATIONS

The system structures considered in the course of this study ranged from retention of a completely manual system exactly like the one currently in operation to the implementation and operation of a totally automated system in which all bibliographic files would be maintained and interrogated by computer. Neither of the two extremes appeared attractive.

A. The Manual System

The continuation of the current manual system, with no change in either input processing or file structure, has a number of serious defects. First, it is evident that some alternative storage medium must be introduced in the near future to alleviate the physical storage problem. Second, the current file structure provides very limited capability for subject control of the material covered making it desirable to develop a file which may be used efficiently for subject searches. Third, although there exists an organization file today, searching this file is a laborious procedure especially when trying to develop a T/O for an organization, a problem which machine search would greatly ease. Thus, the storage problem combined with the desirability of introducing an automatic file maintenance and search capability, creates the need for a new input system designed to capture all data in machine readable form. In addition to alleviating these major problems, the introduction of automation will eliminate the need for hand sorting the 5 x 8 cards and will significantly reduce

sorting and filing errors inherent in a manual system.

B. The Automatic System

The immediate introduction of a completely automated system seems inappropriate in the light of the full DD/I systems study now under way. A study of all of the bibliographic-type projects within the DD/I should be performed before a final system design for any one of them is carried out, since such a study may point out other projects which can be satisfied in terms of the existing bibliographic files or by slightly expanded versions of these files. Requirements for new source coverage may be discovered in the course of such a study. It is also desirable that any files developed be amenable to multiple file query (e.g., that they can be automatically queried based on the results of an earlier search). Therefore, the design of such files should proceed in conjunction with the general file design for the future DD/I information storage and retrieval system. Finally, before converting to a fully automatic system it is desirable to provide for a period of experimentation and adjustment. Particular problem areas requiring study are: subject indexing, query techniques, alternate spellings, depth of coding desired on organization information, and methods to facilitate the man-machine interface, especially techniques which might aid the analyst in the transition from browsing in manual files to obtaining material responsive to his needs by machine interrogation.

Indeed, even if the design and implementation effort were to begin immediately, a parallel interim system would be necessary to cover the time required for full system design. An interim system has the

advantage of being more easily implemented than a final system while maintaining the capability of being converted to a full system at any time.

We therefore conclude that an interim system must be provided which will allow the bibliographic function to continue uninterrupted while providing the basis for the machine system of the future. However, in order to develop the specifications for an interim system it is necessary to develop the basic requirements and the probable configuration of the ultimate system.

The system for storing and retrieving bibliographic information must provide access to this information as a function of personality name, organization name, and subject, or as a function of a logical combination of these. These are the primary retrieval criteria. The other items provided in each entry, such as nationality or journal identification, may also serve as retrieval criteria. The information which may be extracted from the file is a function of the interrogation criteria.

As a function of personality name, one or more of the following may be obtained:

1. Titles of articles and the relationship of the personality to the article.
2. Organization affiliation.
3. Title of personality.
4. Subject information in terms of all keywords describing monographs and articles with which the personality is associated.

5. Other associated personality names.

As a function of organization, one or more of the following may be obtained:

1. Personalities who are affiliated with the organization. Wherever possible, the title of the personality should be provided.
2. Titles of articles published by people associated with the organization.
3. Organization affiliation of co-authors of articles, one of whose authors is connected with the named organization.
4. Keywords defining articles or monographs written by individuals associated with the organization named.

As a function of keywords, or of a logical combination of keywords, the following may be obtained:

1. Personalities associated with monographs and articles in the desired subject area.
2. Titles of articles in the specified subject area.
3. Organization affiliation of personalities associated with articles in the specified subject area.

To provide this capability in an automated system in which the information files are maintained on magnetic tape or on disks, four files--a Master Information File and three Index Files which provide access to the Information File--are planned.

Master Information File--This will be a formatted file containing all of the bibliographic information processed in the system.

Specifically, it will contain an item for each article or monograph covered, where the item will contain all of the information extracted about that article or monograph. The item will have assigned to it a unique accession number by means of which each of the index tapes may reference the item.

Personality Index File--This file will contain an entry for each personality covered in the Master Information File. Each entry will be composed of the personality name and additional identifying information such as professional title, organization affiliation, and nationality code. This will be followed by a list of the accession numbers identifying all articles with which the personality is associated. Attached to each accession number will be a code indicating the type of association (e.g., author, subject of article, or name in text of the article). The file will be ordered alphabetically by personality name.

Organization Index File--This file will contain an entry for each organization covered in the system. The organization will be identified by a code which will indicate the organizational hierarchy and the file will be ordered on this code. Associated with each organization will be a list of accession numbers for monographs or articles which contain references to the organization.

Subject Index File--This file, which will be ordered alphabetically by keyword, will contain each keyword currently in use within the system. Associated with each keyword will be a list of all accession numbers of articles for which the keyword was used as a descriptor.

The files just described must be developed and continuously updated by computer using machine readable input. An automatic search capability must be provided which will be responsive to the information requirements of the bibliographic file users. It is also probable that print-outs of selected information fields from the files will be required periodically to provide desk aids for the analyst. For example, a listing of all personalities and their organization affiliation might be printed annually for hand searching. Two major sub-systems must be developed before the fully automated system can become operational. First, a file maintenance and retrieval program system must be designed and produced which will perform all of the above operations. Second, an input processing system must be developed which will provide data in machine readable form.

C. The Interim System

Although it is felt that the design of the full machine processing system would be premature at this time, it is not too early to begin the collection of data in machine readable form. Basically, an interim system is proposed which would: (a) start capturing bibliographic input for machine processing as soon as possible; (b) use the input to generate the 5 x 8 cards for manual filing into the author and organization files; (c) provide subject control for each entry using a magnetic tape file for storing the data; (d) begin the experimentation with query techniques for forming subject search by computer.

1. Input

The ideal source documents for the intelligence requirements served by the bibliographic files consist of the following:

a. Russian scientific and technical monographs and periodicals.

b. Eastern European scientific and technical monographs and periodicals.

c. Dissertations published in the Soviet Union. (References to these are extracted from Knizhnaya Letopis.)

d. Scientific and technical publications of other Communist countries, especially China.

In the actual system, the degree of coverage on each of the above categories will be based upon processing costs.

It is suggested that, initially, all peripheral input sources be dropped. However, other sources may be exploited in the future to augment the bibliographic files. These include:

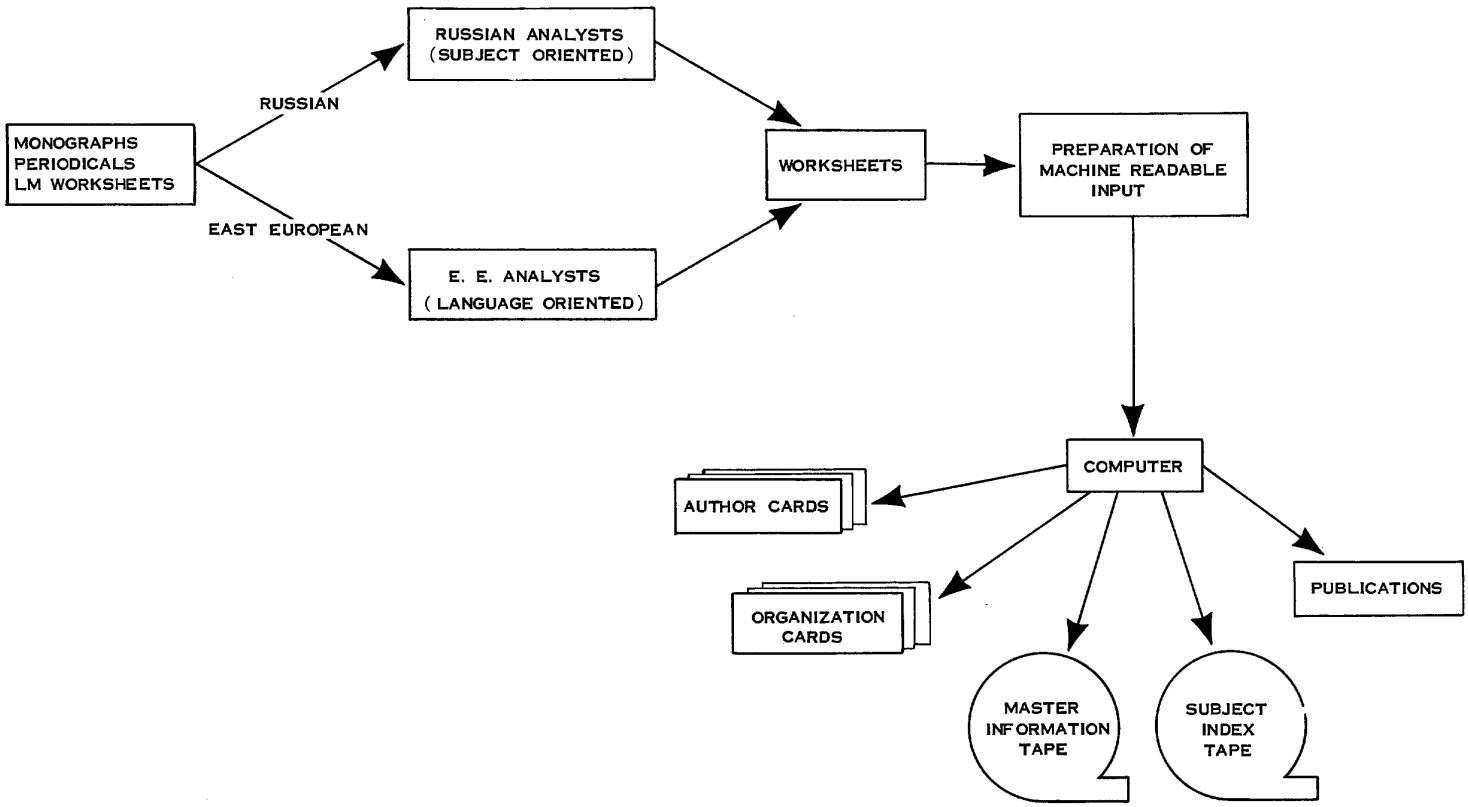
a. STEP cards.

b. Soviet and East European political and scientific cards produced by FDD.

c. Formatted biographic information similar to that maintained in the dossier system.

2. Information Flow

The input processing operation envisaged for the interim system will be, in many ways, similar to the current processing operation. Monographs, periodicals, and NLM worksheets received for processing will be logged in and a check made to ascertain that the item has not been processed earlier. For periodicals, the official journal code will be noted on a routing sheet which, together with the book or journal, will



INFORMATION PROCESSING — INTERIM SYSTEM

FIGURE 14
SECRET

be given to the appropriate translator for processing. For Russian scientific and technical material, the translators will be subject oriented and will concentrate on works covering a limited area of knowledge. The Eastern European material will be processed on a language basis since, for any one language, there is not enough input to warrant a subject breakdown.

The analyst-translator, when he receives a monograph or journal, will fill out the required work sheets. One work sheet will be generated for each monograph and for each article in a periodical. A detailed description of the items of information carried on the work sheet will be given in the next section.

The worksheets will be checked by a senior translator and then routed to the typing section where machine readable text will be produced. One method for producing this text is through the use of Flexowriters producing paper tape much as it is being done in the Cross Check system. However, in view of the fact that complete control over the input is available and because of the slow speed of Flexowriter operations and the availability of optical scanners, consideration will be given to utilizating a character reading device for transforming typed data sheets to machine language.

When the data records enter the computer, they will be checked for validity. Errors will be printed out with an indication of the type of error found. The acceptable records will then be processed to provide 5 x 8 cards for filing in the manual card files, entries will be made on the Subject Index Tape, and the data records will be appended to the

Master Information File for use in extracting information in response to subject searches. The Master Information File will also serve as a permanent file which will form the basis for the fully automated system to be introduced at a later date. Since it is felt that the Master Information File may eventually carry information other than straight bibliographic information extracted from source material, the machine record will include a code indicating the source of the record to allow the analyst to ascertain the validity of the information indexed.

In a later section we will discuss the requirements for published indexes and the procedures to be used for organizing and printing the requisite information.

To produce cards for the manual card files, the information in each entry will be examined to determine the number of cards needed for filing that entry in both the author and the organization files. The record will then be written out on tape that number of times. Those duplicates destined to produce records for the author file will be written on one tape with the name of the author on which the record is to be sorted written first. This tape will then be sorted by author name. Similarly, the duplicate records for the organization file will be written on another tape with the organization which is to serve as the sort key and the personality name associated with that organization written first. These entries will then be sorted by organization code as the major key and by personality name as the minor key.

These two sorted tapes of data will then be processed for printing 5 x 8 cards. The information for each bibliographic record will

be formatted for printing and the results will be produced on continuous form 5 x 8 cards. At the completion of this process the cards, which are already in the proper sort sequence, will be ready for hand filing.

The Master Information Tape will be updated by assigning to each unique entry an accession number and then appending the complete, formatted entry and the accession number to the end of the existing tape file. At the same time, the accession number for the entry will be added to the Subject Index Tape in the record for each keyword associated with the entry.

3. Input Format

Each entry in the bibliographic file will be composed of as many of the following items as are available and appropriate:

a. Translated title of monograph or article. Since translations of a title may vary it is necessary to carry the transliterations of monograph titles to provide for unique identification of the monograph.

b. Year of publication.

c. Periodical identification information.

(1) Numeric code to designate periodical.

(2) Volume number.

(3) Issue number.

(4) Number of first page of article.

d. Subject coding provided through keywords extracted from the title. Supplementary keywords, with a special character to separate the individual keywords, will be added to the entry to further describe the information content when needed.

For each personality of interest who is associated with the monograph or article, the following set of information will be carried in the entry:

a. Name of personality in as complete a form as possible. Names of authors, authors' superiors, editors, and technical editors are of particular interest.

b. The professional title of the individual named, in translated form. For personalities who are heads or deputy heads of the organization with which they are affiliated, a special code will be carried indicating this in the relationship field.

c. A code or set of codes indicating the relationship of the individual named to the article. The relationships of interest are:

(1) Author code to be used for both authors.

(2) Non-author code to be used for authors' superiors and editors.

(3) Deceased personality code.

(4) Biography code for individuals who are the subject of the article.

d. Alphabetic code to indicate organization affiliation. In addition to the code, a transliteration of the organization name as given in the text will be carried in the record if the code is not detailed enough.

e. A nationality code will be provided for each personality carried

in the entry. Eight codes will be used initially to represent the following: USSR, Poland, Czechoslovakia, Hungary, Rumania, Bulgaria, Yugoslavia, All Other. If the nationality of the individual is known and is different from that of the country of origin of the journal in which he is publishing, his true nationality should be coded. Otherwise, the nationality should be coded as the country of origin of the journal. Oriental names are to be coded as "All Other."

4. Interrogating the Files

Interrogation of the author and organization files will continue as it is being done today.

To provide access to the file as a function of subject, two procedures will be used. First, a Keyword-in-Context index will be generated probably monthly with semi-annual and annual compilation of all titles added to the machine file. This will serve as a desk aid to the analyst and will indicate to the analyst the keywords currently on file in the machine system. Simultaneously, a query program will be developed which will permit the retrieval of entries which match a logical combination of request keywords. A basic capability in this area can be provided by permitting retrieval of all entries which contain any one of the keywords contained in the request. The next step in the procedure is to refine this process to eliminate, as far as possible, the false drops which will result from the extremely simple technique described above. Another step is to provide for the retrieval of documents described by keywords synonymous with or similar to those used in the request, thus reducing

the chance of missing a relevant document and, at the same time freeing the analyst of the task of thinking up all keywords which might be relevant to his needs.

The elimination of false drops may be accomplished by providing for a threshold match for retrieval; that is, retrieving only those references which have a predefined percentage of keywords which match those keywords in the text. The development of the threshold value will be determined experimentally.

The synonym or similar word problem may be handled by development and use of a synonym dictionary and a thesaurus, or by use of an association factor such as that proposed by Stiles 1/ to enlarge the list of terms given in a request. Actual experimentation with an operational file will provide the best means of determining the technique or combination of techniques required to solve the problems associated with this particular file.

In addition to improving the retrieval techniques through experimentation, the development of better indexing techniques will be studied. For example, it may be that instead of using keywords which consist of single English words for defining information content, a significant gain in retrieval accuracy may be obtained by using descriptors consisting of groups of related terms. These techniques tend to remove

---

1/ The Association Factor in Information Retrieval, H. Edmund Stiles, Department of Defense, Journal of the Association for Computing Machinery, April 1961, pp. 271-279.

some of the ambiguity inherent in individual words in the English language since they provide, in effect, contextual information for evaluating each word.

PART VII.

OTHER PROBLEM AREAS

A. File Conversion

Upon the introduction of an automatic system in which the data files are stored on magnetic tape or on disks, the question will arise as to the form in which the existing hard copy files are to be maintained. One possibility is to convert the entire existing file to machine readable form and to combine it with the new data as it is produced. In the case of the Bibliographic Project files, this solution seems unreasonable in view of the size of the file. The cost of such conversion in both dollars and time does not seem warranted, especially since the long term value of the hard copy records is questionable. A second possibility is to microfilm the existing file. This, too, is very expensive and does not provide easy access to the file content since microfilm is hard to search (although some of the newer storage media--e.g. Lodestar--might help overcome this difficulty). Again, the question of the long term value of the file negates the value of conversion. Finally, it is possible to leave the file as is or to convert selectively. Possibly, a combination of these is desirable. By leaving the file as it is now and beginning to build a machine file with current information, it will be possible to evaluate the usefulness of the manual file by maintaining statistics on its usage as a function of time. Based upon such a study, the manual file could be gradually abandoned, it could continue in its present form, or criteria for selective conversion could be devised. While it is true that biographic information continues

in importance for a long period of time, subject information may become obsolete as a function of time.

During the time that data are being captured in machine readable form and cards are being added to the file based on these data, it is advisable that the cards so generated be printed on paper which is easily distinguishable so that when the machine system is put into operation the cards which duplicate the machine records may be easily purged from the manual file.

B. Special Dissemination of Data

Within the Agency, special data files are maintained to serve special functions. Information for inclusion in these files may be generated from the bibliographic input and routed to the cognizant office or individual. If the criteria for disseminating the data may be stated in terms of one or more elements of the bibliographic record, the selection and dissemination of pertinent information may be achieved automatically at the time that the data is first introduced into the machine.

For example, biographic information is extracted from the bibliographic input and incorporated in the existing biographic files. Based upon this extraction translation requests are forwarded to FDD. These operations, in the interim system, would be performed as the input is processed initially. Copies of all entries containing the relationship code, BIO, indicating a biography, would be disseminated to BR and FDD for appropriate handling.

To enable BR to remain current on the personalities who are directors

and deputy directors of scientific and technical organizations within the USSR and the Satellite countries, new entries in which personalities are tagged as heads or deputy heads, would be printed for inclusion in a special "director" file. Of course, such a file would not be needed once the fully automatic system is installed since this information could be extracted on demand.

With the introduction of more detailed subject control, and the capability for performing subject searches by computer, it is probable that the bibliographic files would fill the needs of the Biology and Medicine Branch of OSI both of which currently maintain outside contracts to obtain this control. ORR's need for subject control over Vestnik Svyazi would be satisfied also. In addition, support could be provided to the OSI contractors doing studies in the scientific state of the art by notifying them of pertinent articles in the Bloc open literature.

C. Publication of the MIRA

The requirements and design of the proposed bibliographic system are based primarily on a consideration of the system's function as an intelligence tool. It is assumed that the format of and indeed the very existence of the publication Monthly Index of Russian Accessions is a secondary consideration and should not have a strong influence on the design of the bibliographic files.

The fate, however, of the MIRA must be considered in the light of the proposed bibliographic system design. Under the configuration just described, and assuming that LC continues to process its non-scientific and technical material as before, the product resulting from the

processing of the Russian scientific and technical literature bears little resemblance to the output of the non-scientific and technical process. In particular, the proposed design does away with the assignment of LC subject categories to the scientific and technical literature, replacing this type of subject control with a keyword control. This makes it impossible to publish an index which is organized by subject since the subject categorization has been eliminated. We cannot achieve compatibility with non-scientific and technical indexing by using keywords on that material since it is felt that keyword indexing is not suitable for the definition or description of non-scientific and technical material. Hence, to continue publishing a joint index seems a poor choice. It appears that if publication is to continue, two separate indexes must be produced with the non-scientific and technical index continuing in the same form as the current publication. Since the Russian processing is organized on subject lines it would not be difficult to completely separate the two operations.

Concerning ourselves only with the publication of the scientific portion of the journal, the most reasonable approach for disseminating scientific titles for which no subject categorization has been provided, is through a keyword-in-context index. The automatic production of such an index would be a relatively simple matter since the data would all be in machine-readable form. In addition to publishing the monthly MIRA from the machine-readable data, cumulative issues for semi-annual and annual publication could be provided easily.

The MIRA enables the user to identify articles and monographs in

subject areas of interest to him. Keyword-in-context indexes have been used successfully for such a purpose in such publications as Chemical Titles, Chemical Patents, Meteorological Titles, and Biochemical Title Index.

PART VIII.

CIA RELATIONSHIP TO CROSS CHECK

Three major possibilities exist for the configuration of the Bibliographic Project input processing system. In the first case, two completely separate input processing systems could be maintained at the Library of Congress, one sponsored by CIA and the other by FTD, with each system processing all of its own input.

A second possibility is to maintain separate input processing systems, but to process only once the information common to both systems and to share the results. In this instance, the input for both the manual data files and the machine readable subject control tape would be derived from a combination of the Cross Check and the LC products. If Cross Check institutes the changes in format proposed to them and if the LC product is produced based upon the specifications provided in the section on Input Format, the two products will be compatible.

A third alternative would be to establish a joint input processing effort supported by both CIA and FTD under single management at the Library of Congress.

The only advantage to the first possibility lies in the fact that CIA would not have to rely on anyone else as a data source, thus making it easier to introduce changes in the input processing phase during system design and implementation. However, complete duplication of processing is retained. In both the second and third case, duplication of effort is eliminated and hence one of these two systems seems most desirable due to the tremendous cost-saving involved.

Of the latter two courses of action, we favor the combining of the CIA and Air Force programs. All logic dictates that where costs are such a significant factor, every effort should be made to coordinate operations as closely as possible and eliminate any unnecessary overlap and duplication. This could best be done not simply by the establishment of effective liaison and by the precise allocation of processing responsibilities, but by actually placing the Bibliographic Project and Cross Check input activities under single management authority within the Library of Congress. If the MIRA publication itself is dropped as a result of the current survey, the Library of Congress, in all likelihood, will wish to transfer the bibliographic card effort to another Library department anyway. Even if the publication lives on, the Library would probably look with favor on a merger of its two literature-indexing contracts.

Our discussions with the Air Force have led us to believe that they are willing to modify Cross Check to the extent necessary to make a coordinated bibliographic effort possible. We suspect that they would also look with favor on a proposal for administrative centralization of the two activities. In their last communication to us on the subject, dated 15 August 1962, the Air Force Project Monitor of the Air Information Division, George H. Rogge, Jr., stated that: "The changes as proposed appear generally feasible and it remains, then, to phrase the proper methods and instructions for implementation. . . .We shall expect to hear from you when you are ready to finalize the joint procedures."

PART IX.

CONCLUSIONS AND RECOMMENDATIONS

This study was initiated to determine the advantages to be gained from the application of electronic data processing techniques to the problem of bibliographic control of open literature for use by intelligence analysts and, if possible, to develop a coordinated Sovbloc literature exploitation program within the intelligence community.

In considering the most desirable structure for a bibliographic control system our thoughts have ranged from the retention of a completely manual system (like that now in operation in Biographic Register) to the introduction of a fully automated system in which all files would be maintained and interrogated by computer. We have concluded that the immediate introduction of a completely automated system would be inappropriate because:

1. Identification and investigation of all of the bibliographic projects within the DD/I should be completed before attempting to automate any one of them. This would insure that file design, retrieval technique, data coverage, and information extracted would be general enough to satisfy a large number of these projects.
2. The design of the bibliographic machine files should proceed in conjunction with the general file design for the future DD/I information storage and retrieval system.
3. A period of experimentation and adjustment must be provided with emphasis in the areas of index design, development of

query techniques, study of the alternate spelling problem, determination of depth of coding required on organization information, and development of methods to facilitate man-machine interface.

We believe that an interim system should be provided which would allow the bibliographic function to continue uninterrupted while providing for the machine system of the future. Interrogation of the author and organization files would, in the proposed interim system, continue as it is being done today. However, in preparation for the fully automated system, we recommend that the interim system:

1. Start capturing bibliographic input for machine processing as soon as possible;
2. Use the input to generate, by computer, the 5 x 8 cards for manual filing into the author and organization files;
3. Include the design of a subject index;
4. Generate a magnetic tape file of subject index information;
5. Begin experimentation with query techniques for performing subject search by computer.

Our study has also led us to the view that a cooperative effort between CIA and FTD in providing bibliographic control is both feasible and desirable. Cost considerations alone dictate the need to coordinate these two operations in order to eliminate duplication. This could best be accomplished by placing both the Bibliographic Project and Cross Check input activities under single management authority within the Library of Congress.

Looking ahead, we foresee the following to be the major tasks required to implement the proposed interim system:

1. Negotiations must be carried out by OCR with the Library of Congress and the Air Force regarding future management of the Sovbloc literature indexing effort.
2. Discussions must be held with the Air Force to secure final agreement on indexing procedures.
3. A cost experiment must be set up at Library of Congress in which data would be processed according to the specifications for the recommended system to estimate personnel requirements for that system. This will permit a decision to be made on the number of titles that can be covered with the funds available for this project.
4. Statistics should be gathered covering file utilization and specific details on the types of uses and results obtained from the existing bibliographic card files in BR. Such information will be useful for determining the value of the different types of bibliographic files and in the design of the machine files to permit efficient computer processing in the fully automated system.
5. A mnemonic code system must be established for representing, in abbreviated form, the scientific and technical organizations of the USSR and Eastern Europe. BR organization files can most readily provide this information.
6. General system design must be completed including:

a. Documents to be covered.

b. Input format.

c. Method to be used for producing machine readable text.

7. Indexing rules must be specified precisely.

8. An experiment must be designed for testing and evaluating the proposed subject index system.

9. Computer programs must be written.

To begin system implementation we recommend that a full-time study team of three persons be formed. This team should be composed of the following: (a) two system analysts from ADPS; (b) one biographic analyst from BR. In addition, the part-time assistance of a high-level OCR staff member will be required for conducting negotiations with the Library of Congress.

Initially, the team members will concentrate on the design and execution of the input cost experiment. Following the completion of this investigation, they will turn their attention to the final details of the interim system design. Two programmers should be phased into the project at this time to begin preparing the machine instructions.

APPENDIX A.

STATISTICS

By far the most difficult phase of this study has been that of collecting meaningful statistics covering input processing and file utilization which would permit the development of accurate cost estimates of potential future system configurations. Indeed, we have concluded that, given the information available at present, it is not possible to prepare valid estimates. Due to the size of the operations involved, any manipulation of the basic figures obtained serves to magnify the errors which we believe exist in these figures. For this reason, we will limit ourselves to a presentation of the statistics which were collected, develop some figures on the current annual cost to CIA of its various major bibliographic activities, and comment in general terms on the probable manpower requirements of the proposed system. In addition, a controlled experiment will be proposed to obtain the concrete information needed to make accurate cost estimates preparatory to developing a final system design.

1. Input Processing at LC

LC statistics cover processing of East European scientific and technical periodicals and all Russian monographs and periodicals, both scientific and technical and non-scientific and technical. They include the preparation of the Monthly Index of Russian Accessions and of the bibliographic cards for CIA.

FY 62 Processing Figures -- LC

	Monographs	Periodicals	Articles 1/
Russian non-scientific and technical	8,570	1,871	44,928
Russian scientific and technical	9,010	4,890	117,360
EE scientific and technical	None	2/	2/

1/ The number of articles was computed by multiplying the number of periodicals by 24, the LC estimate of articles per periodical in Russian literature.

2/ No accurate figures are available for the number of East European scientific and technical periodicals processed throughout the fiscal year due to the elimination of the East European Accessions Index in December 1961 and the addition of a large number of periodicals through the second half of fiscal 1962. However, from January 1962 through July 1962, the Library of Congress estimates that 26,000 articles from 2,450 journals were processed.

The Russian non-scientific and technical material was processed for inclusion in the MIRA only; the Russian scientific and technical material for the MIRA and the bibliographic files; the EE scientific and technical material for the bibliographic files only. Based on the processing figures it is estimated that 126,370 Russian scientific and technical entries were covered for the bibliographic files. In light of the fact that BR estimates that there are an average of 3.2 personality names per entry, and that 180,000 cards (or approximately 56,000 entries) are filed in the author file each year covering both Russian and EE material, the validity of the LC processing figures is questionable.

The T/O for the Library of Congress activity totals 70 slots with 42 charged to the Bibliographic Project and 38 to the MIRA. The average grade of the slots charged to CIA is 7.4; the average grade in the MIRA slots is 5.8.

In addition to processing figures, the Cyrillic Bibliographic Project Annual Report for 1961/62 lists the following backlogs as of July 1, 1962:

Typing	36,338 entries representing 2,527 periodical issues. (This backlog is increasing.)
Translating	735 periodical issues. (This is a reduction of 165 issues over the year resulting from authorized overtime.)

2. CIA Processing

When the 5 x 8 cards from LC arrive at CIA, they are screened, distributed, sorted and filed. Of the cards which are received a very large number of the duplicate copies are discarded.

Personnel--In BR's Support Branch three people are responsible for

screening the cards, routing the Russian organization cards to the USSR Section, routing the Satellite organization cards to the Satellite Section, sending the author cards (appropriately underscored) to the pool for sorting, reviewing and correcting sequencing errors in the cards returned from the pool, filing the sorted author cards, responding to requests, and refiling. It is estimated that two additional people are required by the Support Branch to adequately handle the job. Pool sorting time averages 1,300 man-hours per quarter. One person is responsible for maintaining the USSR organization file; 80 hours per month or 1/2 person is needed to maintain the Satellite organization file.

Initial Filing--Approximately 5,000 cards are filed in the USSR organization file per month. They are generally screened before filing so that only one card is kept in the file for each person in an organization. Approximately 4,000 cards are filed per month in the Satellite organization file. Much less screening is done on these cards, hence they more nearly represent the actual number of cards routed from the Support Branch. Approximately 25,000 author cards are filed per month. These include 15,000 cards received from LC, 8,000 STEP cards, and 2,000 miscellaneous cards generated within CIA. Since STEP cards normally replace LC cards already on file, these are not true additions to the file but rather replacements.

Queries--Estimates of file utilization are totally unreliable. Formal use of the bibliographic files resulting in the reproduction of cards for a requester may be ascertained by examining the request

log. Such a survey indicates that 150 formal requests involving 2,400 names are levied against the file per year, resulting in the reproduction of 45,000 cards. However, the major use of the file is not shown explicitly in the request log since it involves examination of the file but no reproduction of file content. The files are used to answer outside requests either through written reports, oral responses, or inspection of the file. It is estimated that 8,000 author searches and 800 USSR organization searches of this type are performed annually. This figure was arrived at by reading the request sheets for the 3-month period, April-June 1962, and counting the number of names for which information was required in which it is probable that the bibliographic files were searched but where no reproduction of the bibliographic cards was noted.

The bibliographic files are also used in the course of internal processing in BR. Estimates of such utilization are impossible to obtain except by actually keeping count of analyst use of the file over a reasonable time period. An attempt was made to arrive at this figure by determining the number of cards refiled and estimating the cards pulled per name. However, no adequate refile figure could be obtained and, more important, many searches do not involve removal of material from the file.

It is recommended that a count be kept on all file utilization for a one-month period to determine for each file such information as:

a. Number of queries.

b. Number of names involved.

c. Type of information extracted.

d. Number and types of unsuccessful queries.

e. Requirements on response time.

Subject searches are handled today primarily by machine runs on the Who's Who cards with search of the bibliographic file only as a last resort because there is no actual subject control to assist the searcher. The number of such requests in increasing. It is the opinion of BR analysts that they would be greatly aided in providing responses if automatic subject search were possible but no accurate information can be obtained beyond this generalization.

File Holdings--The following are estimates of current file holdings:

Author File	2,000,000 cards (500,000 discrete names)
USSR Organization File	250,000 cards (3,400 discrete organizations)
Satellite Organization File	46,000 cards

3. Cross Check

Cross Check, in 1961, processed 150,000 entries where an entry consists of a set of information covering one personality name or one organization. Of these, 75,000 had been typed on Flexowriters by the end of the year. They estimate that 10% of their names are extracted from newspapers; the remainder result from processing scientific and technical periodicals for which they provide the following statistics:

Average number of issues per periodical---------10

Average number of documents per issue----------18.5

Average number of entries per document----------4.2

Their statistics covering processing rates are:

Average 5.53 entries per hour per person completed by professional personnel (including translation of titles, selection of uniterms, filling-out of the work-sheet, final reviewing for accuracy, etc.).

They have a total of 17 professional personnel.

Average 15 entries per hour completed by sub-professional personnel (including searching, operating of Flexowriters, and other related clerical duties).

They have a total of 8 sub-professional personnel.

All of Cross Check's estimates are based on a 210-day work year.

Evaluation of these statistics indicates that the processing rate (not including Flexowriting) is 1.3 articles per hour or one journal issue every 14,2 hours. These figures appear extremely low. To provide a meaningful figure for use in estimating personnel requirements for the interim system proposed here, it is recommended that an experiment be set up at LC to process a set of data based on the plan for the interim system. It is suggested that EE periodicals provide the test base since no conflict then arises with the MIRA effort while at the same time they form a representative sample. In the course of this experiment translation rates must be established. Flexowriter processing rates should be obtained including study of the error correction problem in paper tape. At the same time more detailed

consideration can be given to the applicability of optical scanners.

4. The Proposed System

Under the new system, excluding the possibility of cooperating with Cross Check, the number of titles that will be processed for the East European and Russian scientific and technical literature will be about the same as the number processed now. With processing of the basic entry in the proposed system not significantly more complicated than under today's system, and with the introduction of a single processing pass (MIRA input would be a by-product of the bibliographic file processing), the number of translators required should not increase. If it is possible to make use of Cross Check's product, the number of titles remaining will be reduced thus easing the translation processing load.

The number of personnel required for providing machine-readable text will depend, obviously, upon the method to be used for the conversion. If a character reader is used, the number will be much less than is currently required to type the MIRA cards and the bibliographic mats.

Since all sorting will be done by machine, the need for personnel to perform this service will be eliminated. In addition, the quality of the sorting will be much better than that currently available thus simplifying the job of filing the cards. During the course of the interim system, the initial filing, file search, and refile problems will remain essentially the same.

5. Annual Cost of Major CIA Bibliographic Projects

The following figures represent current operating costs and are considered accurate. However, it should be noted that the personnel costs covered by these figures are not adequate for the job. As a consequence, backlogs exist in almost all operations as noted above.

Bibliographic Project	
Sorting (CIA pool)	$10,600
File Maintenance (BR)	19,400
OCR/LC Contract	336,380
OSI/Medical Contract	27,000
OSI/Agriculture Contract	70,000
Total	$463,380

APPENDIX B.

PROPOSED RUSSIAN TRANSLITERATION SYSTEM

A transliteration system from the Cyrillic to the Latin alphabet has been developed which is a slight modification of the BGN system.

А	A	Л	L	Ц	TS
Б	B	М	M	Ч	CH
В	V	Н	N	Ш	SH 3/
Г	G	О	O	Щ	SHCH
Д	D	П	P	Ы	Y*
Е	E or YE 1/	Р	R	Э	E
Ж	ZH	С	S	Ю	YU
З	Z	Т	T 2/	Я	YA
И	I	У	U	Ъ	"
Й	Y	Ф	F	Ь	'
К	K	Х	KH		

The changes which have been made to the BGN system permit the transliterated text to be rewritten in the Cyrillic with no ambiguity. Since no letter in Cyrillic is transliterated as the letter H, the combinations ZH, CH, and SH do not result in any ambiguity. The letter combinations YU and YA, if used to represent the Cyrillic Ю and Я, cannot be confused with the transliterations of ЫУ and ЫА since Ы is transliterated as Y*. To differentiate between the transliterations of Щ and ШЧ, the symbols ШЧ when appearing together will be represented

---

1/ Transliterated as YE when starting a word or after vowels (A, E, И, О, У, Ы, Э, Ю, Я) and after Ъ, Ь.

2/ T will be transliterated as T* when the following letter is a Cyrillic С.

3/ Ш will be transliterated as SH* when the following letter is a Cyrillic Ч.

by SH*CH. Similarly, Ц and TC are differentiated since TC is transliterated as T*S. Grammatical rules enable one to differentiate between the transliteration of the Cyrillic E and Э. The Э is used only to start a word; it is transliterated as E. When the Cyrillic E starts a word, it is transliterated as YE.

The *, ', and " will each be treated as a separate character within a transliterated word and hence will affect the computer sequencing of words. The characters blank, *, ', and " sort in the order shown and are all less in value than the letters of the alphabet.